Genetic variations in connection

VALORIZATION

From researchers to researchers

Nowadays, the internet provides an enormous number of tools that scientists can use to analyze, consult and store data obtained from the wet laboratory. However, the time has not yet come where ”one tool does everything”. Thus, often researchers spend time to search for tools that potentially provide the service required, check if and how the tool performs the tasks desired, investigate if the way to run it fits with the user competencies and only then are able to test that analysis tool for his/her research. These operations require investment of time, effort and expertise that often biologists do not have. For this reason, the review article presented in this thesis is a valuable resource to those researchers who wish to interpret genetic variants using the power of the biological pathway knowledge. The fact that the applications reviewed have the peculiarity of providing visualization features and a user interface makes the tool evaluation even more appealing for those researchers that are not very familiar with programmatic tools. Moreover, the relevance of such a tool inventory is not only for those that will use the tools, but also for bioinformaticians or computational biologists that develop such applications. These researchers need the opinions of the users and experts of the field, in order to understand if the tools work properly, and what type of improvements are needed to perfect them.

Another effort in raising awareness on using bioinformatics methodologies for data interpretation is presented in Chapter 5, where we make publicly available a video tutorial and a web-session of the genetic networks obtained by combining multiple data types. Sharing with the scientific community such resources is important to: i) encourage the application and validation of the data integration method in other datasets, ii) reproduce the analysis step-by-step in an easy to follow manner, iii) further investigate the results with different perspectives.

Finally, presenting an inventory and evaluation of tools, sharing data and explaining the methodology are all efforts that improve communication between researchers. This is a key aspect in modern science, where due to the technological advancements expertise becomes more and more specialized, but at the same time where interdisciplinary skills are an essential requirement.

Towards an improvement of data interoperability

An important concept in data science is to enable computer systems and software to exchange and easily make use of information, this concept is also known as interoperability. This is an essential aspect for improving the understanding of data in life sciences. Nowadays, it is clear to every researcher that the advancement of the new technologies create the ”big data” issue. This is not only a problem related to data

storage, but also to data use and interpretation. In this regard, putting the effort of increasing and perfecting data interoperability worldwide contributes to building a better structure of the body of human knowledge. Moreover, data analysis tools can see improved performances if the data are properly linked, and researchers smoothly can run workflows that comprehend the usage of several tools and environments. Such implementation results in saving time and money for complex and specialized data analysis. In line with this value, Chapter 3 includes an implementation of a mapping protocol for existing identifiers called BridgeDb, which is an existing resource that makes data interoperable because it matches the different types of identifiers related to the same biological entity. The map already stores genes, proteins and metabolites of different species including human. The implementation described here is a gene-to-variant and variant-to-gene mapping that adds a new dimension in the database related to the molecular world/dogma. The tangible benefit of such an implementation is evident in the application of the mapping database to the tools that analyze biological data. Indeed, the map can be used in pathway and network applications (e.g. PathVisio and Cytoscape) that allow analysis and visualization of different data types.

Looking forward to the future of precision medicine

SNPs are the common genetic variations in the human genome, and currently there is growing business activity around the concept of precision medicine, much of which relies on the possibility to perform genetic tests to predict diseases or improve physical or health conditions. Numerous companies, rather than hospitals or clinics, are providing genetic tests for several types of purposes. In those tests specific variants are assessed and related to risk of certain diseases, food intolerances, or even improvement in physical performance. For this reason, tools and methods that are able to support and improve the interpretation of the biological effects of the genetic variant are in high demand. In this thesis, a workflow is presented that combines multiple data types primarily in order to understand the effect of the genetic variant in specific clinical conditions, namely T2DM and obesity. Moreover, genetic reference networks of SNPs associated with obesity are proposed, in an attempt to provide a visual instrument to elucidate the biological and medical fucntion of the variants. These maps of SNPs, genes, pathways, and their relationships to each other can be used in different ways by different stakeholders who are interested in obesity and personalized treatments. Experts in the field of obesity can explore the networks to generate novel hypotheses or confirm results related to the functional role of BMI SNPs and their possible effects on gene regulation by influencing epigenetic marks. Clinicians involved in precision medicine also can benefit from such networks. For patients with available SNP genotyping data, the health care team, in theory, can determine susceptibility to certain diseases by consulting the reference

networks. Exploring those SNPs present in the network and the patients genotype can assist interpretation of the impact of the patient's alleles, linking them to the gene and the functional context in which they are involved. For example, the occurrence of several genotyped SNPs from the patient that indicate presence of risk or effect alleles that occur in the same or related pathways, can prompt the health care team to evaluate if those processes, in relation to the specific tissue, are relevant to the patient's current or future condition.

Conclusion

Current fields of bioinformatics and systems biology have developed new technologies and methodologies to further explore life science data. However, often the technical specialization within this field increase the gap of biological understanding, mostly due to a jargon issue. This is the reason why in the Bioinformatics and System biology communities despite the technical advances, researchers need to be able to be good translators within the biological and computer science area of knowledge. This role itself has a strategic influence in terms of societal and economic value, and this thesis contributes to highlight it. The societal value of the work reported relate to the aspect of improving communication within researchers of the same discipline (like bioinformatics), but also stakeholders of different fields, (e.g. bioinformaticians, biologists and clinicians). On the other side, the economic value is less tangible to the public because the advancements presented, such as the workflow for data integration and interpretation and an example of improvement in tool interoperability, directly benefit the area of basic research. Improved methods to perform analysis and interpret results is a key to innovation, without wasting public funds. The work performed in this thesis definitely contributes to this purpose.